Species-Specific Multi-Stage Clustering for Animal Re-ID

Bo-An Chen[1,†], Pin-Hao Shen[1,†] and Yi-Yu Hsu[1,*]

[1]Miin Wu School of Computing, National Cheng Kung University, No. 1, University Road, Tainan City 70101, Taiwan (R.O.C.)

Abstract

Individual animal re-identification (Animal Re-ID) aims to distinguish individuals within the same species from images, supporting non-invasive wildlife monitoring without capturing, tagging, or disturbing animals. AnimalCLEF 2026 formulates this problem as identity discovery and clustering, where each test image must be assigned a cluster label and each cluster should correspond to one individual animal. This setting is challenging because the species subsets differ in available training labels, visual cues, and dominant clustering errors. Some subsets mainly suffer from fragmented clusters caused by viewpoint or pose changes, while others have few images per individual or no labeled training set. We propose a two-phase, multi-stage global-local clustering pipeline (Stages A–F) for this multi-species setting. The first phase builds a shared clustering foundation through Stage A, which selects species-wise global representations, constructs pairwise similarity matrices, and applies HDBSCAN to obtain initial clusters. When training labels and local matching evidence are available, Stage A combines calibrated global similarity with SIFT + LightGlue match counts. For Texas horned lizard, which has no labeled training set, it uses an uncalibrated global fallback plus a conservative cluster-level local check. The second phase applies species-specific correction through Stage B to Stage F rather than using a universal post-processing rule, including orientation-aware merging for Lynx, belly-region graph replacement for Texas horned lizard, region-normalized merging for Salamander, and signed relation repair for SeaTurtle. The submitted release achieves 0.74911 Public Leaderboard ARI and 0.67537 Private Leaderboard ARI, ranking fourth in the AnimalCLEF 2026 competition under the team name HSU Lab. These results suggest that shared initial clustering followed by species-specific correction is a practical design for multi-species identity discovery and clustering. Our code is publicly available on GitHub at https://github.com/hsucomputinglab/animalclef2026-species-specific-multistage-clustering.

Keywords

Animal re-identification, identity discovery, clustering, local feature matching, AnimalCLEF 2026

1. Introduction

Image-based wildlife monitoring provides a non-invasive way to observe animal populations over long periods of time. By analyzing images collected from camera traps, field surveys, or citizen-science platforms, researchers can monitor repeated appearances of individual animals without capture, tagging, or direct disturbance. Individual animal re-identification (Animal Re-ID) is a key requirement in this setting because it distinguishes individuals within the same species rather than only recognizing the species category. The AnimalCLEF 2026 task is organized within the LifeCLEF 2026 lab and focuses on discovery and re-identification of individual animals [1, 2].

Animal Re-ID is challenging because identity cues are often subtle, local, and sensitive to imaging conditions. Useful cues may appear as spots, stripes, facial scale patterns, skin patterns, or other natural markings, but these cues can be affected by pose, illumination, occlusion, viewpoint, background clutter, age-related changes, injuries, and sensor or habitat domain shifts [2, 3, 4, 5, 6]. At the same time, different individuals from the same species may share similar overall appearance or local markings. A reliable method must therefore preserve fine-grained identity cues while reducing shortcuts based on background, coarse appearance, or acquisition conditions.

Recent AnimalCLEF and Animal Re-ID systems often combine global descriptors, local feature matching, calibration, and species-aware refinement [7, 8, 9, 10]. Global feature backbones produce

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

kevink0725n@gmail.com (B. Chen); chenpinhao910827@gmail.com (P. Shen); yiyuhsu@gs.ncku.edu.tw (Y. Hsu)

0009-0002-2195-3786 (B. Chen); 0009-0002-0382-1392 (P. Shen); 0000-0003-4286-3181 (Y. Hsu)

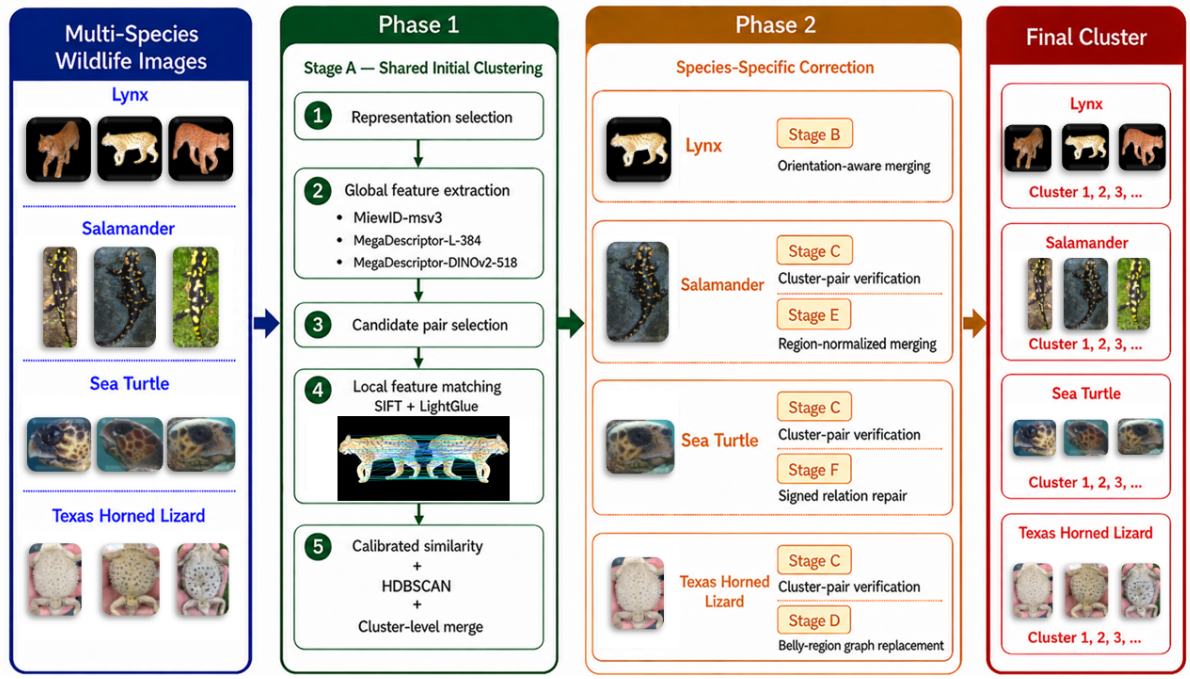

Figure 1: Overview of the proposed two-phase pipeline. Phase 1 builds the shared initial clustering through Stage A by selecting a species-wise representation, extracting global features, selecting candidate pairs, adding SIFT + LightGlue local match counts where applicable, and applying HDBSCAN followed by a conservative cluster-level merge. For labeled subsets, calibrated similarity is used before HDBSCAN, while Texas horned lizard follows the uncalibrated fallback described in Stage A. Phase 2 routes each species through its correction stages: Lynx through Stage B, Salamander through Stages C and E, SeaTurtle through Stages C and F, and Texas horned lizard through Stages C and D.

image-level embeddings that capture overall appearance, color distribution, body shape, and large-scale texture. However, these embeddings may be affected by pose, cropping, and background variation. Local feature matching provides complementary local texture evidence by comparing keypoints or local regions between image pairs [11, 12]. In a multi-species setting, however, the reliability of global and local cues can vary across species, so a single matching rule, merge criterion, or universal threshold may not transfer uniformly.

AnimalCLEF 2026 formulates Animal Re-ID as identity discovery and clustering. This differs from the AnimalCLEF 2025 open-set individual identification setting, where systems mainly made known-individual or new-individual decisions and predicted identities for query images. In the 2026 setting, each test image must be assigned a cluster label, and each cluster should correspond to one individual animal. The evaluation metric is Adjusted Rand Index (ARI), which measures pairwise consistency between the predicted clustering and the ground-truth clustering. The main methodological challenge is therefore not only to estimate visual similarity, but also to decide which same-individual relations should be kept, added, or rejected in the full test set.

This paper addresses the AnimalCLEF 2026 challenge with a two-phase, multi-stage pipeline, illustrated in Figure 1. In this paper, a phase is a conceptual grouping of the method into a shared clustering-foundation phase and a species-specific correction phase, whereas a stage is a single operational step, from Stage A to Stage F, that updates the image-to-cluster-label mapping. The first phase builds a shared clustering foundation through Stage A. It uses selected global representations to construct pairwise similarity matrices, applies HDBSCAN to obtain initial clusters, and, when available, uses local feature matching and score calibration. The second phase applies species-specific correction through Stage B to Stage F according to the dominant errors and data conditions of each species subset.

The contributions of this paper are threefold. First, we present a task-oriented clustering framework that separates a shared clustering-foundation phase from a species-specific correction phase for

the AnimalCLEF 2026 identity discovery and clustering task. Second, we design species-specific correction strategies that address different failure modes across species subsets, including viewpoint-induced fragmentation, conservative merging under few-shot identity distributions, mixed relation-level errors, and the absence of labeled training data. Third, we report submitted development-path results that analyze how the clustering foundation and later correction stages change the Public Leaderboard ARI and cluster-count structure.

2. AnimalCLEF 2026 Challenge

This section summarizes the task characteristics that motivate the proposed pipeline. We first describe the AnimalCLEF 2026 objective and then compare it with the AnimalCLEF 2025 open-set individual identification setting. We next define the clustering error terms used throughout this paper from the perspective of merge decisions, and summarize the species-level data conditions and failure modes that motivate the later species-specific correction stages.

2.1. Challenge objective

The main objective of AnimalCLEF 2026 is Discovery and Re-Identification of Individual Animals [2]. Participants are required to create clusters of images, where each cluster should correspond to one individual animal. The output is therefore not a known class label and not a decision between a known individual and a new individual, but a clustering result over the full test set. In this setting, assigning two images to the same cluster represents a same-individual relation, while assigning them to different clusters represents a different-individual relation.

The task further inherits common Animal Re-ID difficulties, but the four subsets emphasize different ones: viewpoint and orientation change (Lynx), very few labeled images per individual (Salamander), long-term appearance change across acquisition years (SeaTurtle), and the absence of any labeled training set (Texas horned lizard). These factors can make images of the same individual appear different, while making different individuals appear visually similar. As a result, the method must not only estimate visual similarity between image pairs, but also decide which same-individual relations should be preserved or corrected when forming clusters over the full test set.

2.2. Evolution from AnimalCLEF 2025 to AnimalCLEF 2026

AnimalCLEF 2025 was formulated as an open-set individual identification task with database and query images. Each query image had to be assigned either to a known individual in the database or to a new individual, and the final ranking score combined BAKS and BAUS through their geometric mean [3]. This setting favored retrieval, local verification, score aggregation, and novelty detection, because the central decision was whether a query image matched an existing identity or should be rejected as new.

AnimalCLEF 2026 changes the focus from query-level open-set identification to test-set-level identity discovery and clustering. Participants assign cluster labels to all test images, so the system must construct a coherent clustering result over the full test set rather than only rank database candidates or apply a novelty threshold to each query image. This shift makes similarity graph construction, density-based clustering, and cluster-level correction more central to method design. Table 1 summarizes the main differences between the two challenge settings.

Table 1
Task-setting shift from AnimalCLEF 2025 open-set identification to AnimalCLEF 2026 identity discovery and clustering.

Dimension	AnimalCLEF 2025	AnimalCLEF 2026
Main setting	Open-set individual identification	Identity discovery and clustering
Input	database and query images	Unlabeled test images to be clustered
Output	Known-individual label or new-individual decision	Cluster label for each test image
Main decision	Decide whether a query image belongs to a known individual	Decide which images should belong to the same individual
Evaluation metric	BAKS, BAUS, and their geometric mean	Adjusted Rand Index (ARI)

2.3. Evaluation metric and clustering error types

ARI is a chance-adjusted clustering metric that measures pairwise consistency between the predicted clustering and the ground-truth clustering [13]. Formally, given n test images, the predicted clustering and the ground-truth clustering form a contingency table in which n_{ij} is the number of images assigned to both predicted cluster i and ground-truth cluster j, with row sums $a_i = \sum_j n_{ij}$ and column sums $b_j = \sum_i n_{ij}$. ARI is then defined as

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}},$$

where $\binom{\cdot}{2}$ counts the image pairs inside a group. The numerator measures the agreeing same-cluster image pairs corrected for chance, and the denominator normalizes by the maximum achievable agreement. A perfect clustering has an ARI of 1, while a result close to random labeling usually has an ARI near 0. Because ARI is computed from image-pair assignments, it decreases when images of the same individual remain separated or when images of different individuals are grouped together.

In the rest of this paper, we use the terms *under-merge* and *over-merge* from the perspective of merge decisions, rather than as direct replacements for the official over-clustering and under-clustering terminology. Under-merge means that images of the same individual remain split into multiple clusters, while over-merge means that images of different individuals are incorrectly assigned to the same cluster. These errors require opposite correction directions. Under-merge requires recovering reliable same-individual relations, whereas over-merge requires preventing or removing unreliable same-individual relations. Table 2 summarizes the terminology used in this paper.

Table 2
Merge-decision error terminology used in this paper.

Term	Meaning in this paper	Methodological implication
Under-merge	Images of the same individual remain split into multiple clusters	Reliable same-individual relations need to be recovered
Over-merge	Images of different individuals are incorrectly assigned to the same cluster	Unreliable same-individual relations need to be prevented or removed

2.4. Dataset characteristics and species-specific difficulties

These species names and image counts follow the Kaggle competition data release used for our submitted development path [14]. They differ in test-set size, available training labels, visual cues, and

dominant failure modes. These differences directly affect method design because a merge rule that improves one subset may introduce incorrect merges or unnecessary fragmentation in another. For SeaTurtle, the training statistics follow SeaTurtleID2022, a long-span sea turtle Re-ID dataset with 8,729 photographs of 438 individuals [15]. For Texas horned lizard, the dataset motivation is consistent with prior evidence that individual Texas horned lizards can be identified using ventral spot patterns [16]. For Lynx, the images are drawn from CzechLynx, a large-scale camera-trap dataset for individual identification of the Eurasian lynx [17]. The subsets also differ sharply in identity density on the labeled training split: Lynx averages many images per individual (median 17), whereas Salamander is strongly few-shot (median 1, with 53% of individuals appearing only once), and SeaTurtle additionally spans about 14 years of acquisition. Table 3 summarizes the data conditions and correction implications used in this study.

Table 3
Species-level data conditions and the correction strategy motivated by each subset. Test counts follow the Kaggle data release, and SeaTurtle training statistics follow SeaTurtleID2022. Per-individual statistics are computed on the labeled training split.

Species subset	Training data	Test images	Main difficulty and correction
Lynx	2,957 images 77 IDs	946	Images are abundant per individual (median 17 in training) but split across left, right, front, and back views, so orientation differences fragment same-individual images and motivate orientation-aware merging.
Salamander	1,388 images 587 IDs	689	Few images per individual (median 1 in training; 53% singletons) make permissive merging risky and motivate conservative region-normalized merging.
SeaTurtle	8,729 images 438 IDs	500	A 14-year acquisition span drives long-term appearance change, so fragmented clusters and incorrect merges can coexist, motivating signed relation repair.
Texas horned lizard	No labels	274	Calibration is difficult, and reliable cues are concentrated in ventral texture, motivating belly-region graph replacement.

The table highlights why the correction phase must be species-specific. Lynx mainly requires recovery from orientation-induced fragmentation, Salamander requires conservative merging under a few-shot identity distribution, SeaTurtle requires relation-level correction in both directions, and Texas horned lizard requires a different path because no labeled training set is available. These observations motivate a two-phase, multi-stage pipeline. The first phase builds a shared clustering foundation through Stage A, and the second phase applies species-specific correction through Stage B to Stage F rather than using one universal merge criterion.

2.5. Implications for the proposed pipeline

The task setting, ARI metric, and species-level differences show that AnimalCLEF 2026 is not well served by a single clustering rule, universal threshold, or post-processing rule. A permissive merge rule can recover more same-individual relations but may introduce over-merge, while an overly conservative rule can reduce incorrect merges but may leave images of the same individual split across fragmented clusters. Since ARI evaluates pairwise consistency, both risks must be considered together.

The proposed pipeline therefore separates the problem into two conceptual phases. The first phase builds the initial clustering for each species through Stage A, using selected global representations, pairwise similarity construction, HDBSCAN, and local feature matching when suitable evidence is available. The second phase then applies species-specific correction through Stage B to Stage F according to the data condition and main difficulty of each species subset. This design treats Stage A as a

shared clustering foundation and leaves Stage B to Stage F to correct species-dependent errors, rather than forcing all subsets to follow the same merge criterion.

3. Methodology

This section describes the proposed two-phase, multi-stage pipeline for AnimalCLEF 2026 identity discovery and clustering. Here, *phase* refers to the conceptual organization of the method, with a shared clustering-foundation phase followed by a species-specific correction phase. In contrast, *stage* refers to the operational steps executed in the pipeline, namely Stage A to Stage F. The output of each stage is a mapping from image IDs to cluster labels. Stage A creates the initial mapping for all four species and serves as the shared clustering foundation. Stage B to Stage F then selectively update this mapping by merging clusters, replacing a species-specific graph, or modifying pairwise same-individual relations.

The pipeline is executed in the order Stage A → Stage B → Stage C → Stage D → Stage E → Stage F. Stage A produces the initial clustering for every species. Later stages act as species-specific routes. Each stage changes only the species it is designed for, and all other species are carried forward unchanged. Figure 1 gives an overview of this design, and Table 4 summarizes this routing map after Stage A. The effects of these stages are analyzed in Sec. 4.

Table 4
Species-specific correction routes after Stage A. If a species is not assigned to a stage, its cluster-label mapping is carried forward unchanged.

Species subset	Correction stages	Correction role
Lynx	Stage B	orientation-aware merging
Salamander	Stage C + Stage E	cluster-pair verification and conservative region-normalized merging
SeaTurtle	Stage C + Stage F	cluster-pair verification and signed relation repair
Texas horned lizard	Stage C + Stage D	cluster-pair verification and belly-region graph replacement

3.1. Stage A: Global feature extraction and initial clustering

Stage A builds the initial clustering for each species. Its input is the set of test images for each species subset, and its output is an initial mapping from image IDs to cluster labels. The goal is not to solve all species-specific errors directly, but to provide a stable clustering foundation for later correction.

Stage A has four components. It first selects a species-wise global representation from frozen and fine-tuned candidates evaluated during development. It then extracts image-level embeddings and constructs a pairwise similarity matrix for each species. Next, it applies HDBSCAN to obtain initial clusters. Finally, it performs one conservative cluster-level local merge. For species with training labels, the similarity matrix combines calibrated global similarity with SIFT + LightGlue match counts on globally selected candidate image pairs. Texas horned lizard follows a label-free fallback path because no labeled training set is available. Supervised fine-tuning, train-label-based calibration, and train-label-based representation or parameter selection are not used for this subset. Its HDBSCAN similarity matrix is based on the uncalibrated average of MiewID-msv3 and MegaDescriptor-DINOv2 cosine similarities, followed by a conservative cluster-level SIFT + LightGlue check.

3.1.1. Global feature backbone extraction

Stage A uses three global feature backbones to extract an image-level embedding for each image, namely MiewID-msv3, MegaDescriptor-L-384, and MegaDescriptor-DINOv2-518 [4, 5, 18, 19, 20, 21]. These embeddings provide the shared global visual representation used by the initial clustering stage.

They capture broad appearance cues such as body shape, color distribution, pose, and large-scale texture, which are useful for selecting candidate neighbors and forming the initial species-wise clusters.

The three backbones are selected because they provide complementary global feature representations. MiewID-msv3 is a multi-species wildlife Re-ID feature extractor based on EfficientNetV2-RW-M. MegaDescriptor-L-384 is a Swin-L-based animal Re-ID image feature model. MegaDescriptor-DINOv2-518 is a DINOv2-based animal Re-ID image feature model. In our implementation, the extracted embedding dimensions are 2152, 1536, and 1024 for MiewID-msv3, MegaDescriptor-L-384, and MegaDescriptor-DINOv2-518, respectively. For each image, Stage A extracts embeddings from all three backbones. Because these models differ in architecture and pretraining data, using them together reduces the risk that the initial clustering depends too strongly on one model's preference for a particular species, pose, or background pattern. Table 5 summarizes their basic specifications.

Table 5
Global feature backbones used to form image-level embeddings in Stage A.

Backbone	Model family	Image size	Params (M)
MiewID-msv3	EfficientNetV2-RW-M	512 × 512	51.11
MegaDescriptor-L-384	Swin-L	384 × 384	228.8
MegaDescriptor-DINOv2-518	DINOv2 ViT-L/14 variant	518 × 518	304.4

3.1.2. Local feature matching and matcher selection

Global feature backbones capture whole-image appearance, but many Animal Re-ID cues are local. For example, SeaTurtle facial scale patterns [22], Salamander yellow patches, and Lynx side spots may be diluted in an image-level embedding, especially when pose, background, or cropping variation is large. For the species where local matching is used in Stage A, the pipeline therefore compares local keypoints between candidate image pairs [7, 11, 12]. The match count is the number of successfully matched keypoints between two images. A higher match count indicates stronger local texture agreement and is used as complementary evidence for a possible same-individual relation.

To select the local matcher used in Stage A, we compared five matching pipelines under the same submitted development path. As shown in Table 6, SIFT + LightGlue achieved the highest Public Leaderboard ARI among the tested matchers and is therefore used as the Stage A local matcher when local matching is enabled. This result should be interpreted as a development-path comparison rather than a general claim that one matcher is universally superior. In this multi-species setting, local identity cues include scales, fur patterns, spots, yellow patches, and facial scale patterns. Under this setting, SIFT + LightGlue appeared more stable than the tested learned local feature pipelines and the dense matcher in the submitted Stage A pipeline.

Table 6
Stage A local-matcher comparison on Public Leaderboard ARI under the submitted development path. These scores are measured at the local-matcher selection step, before per-species fine-tuned representations and the cluster-level merge; Table 14 shows how the remaining Stage A components reach 0.58280.

Matcher	ARI
SIFT + LightGlue	0.45294
DISK + LightGlue	0.42256
ALIKED + LightGlue	0.40178
SuperPoint + LightGlue	0.38219
LoFTR	0.26408

In implementation, Stage A first detects SIFT keypoints for each image and then uses LightGlue to match candidate image pairs. To reduce computation cost, Stage A does not perform full pairwise SIFT +

LightGlue matching for all images within a species. Instead, it uses calibrated global similarity to select the top 300 candidate images for each test image, as described in Sec. 3.1.4, and applies LightGlue only to these candidate image pairs. The resulting match count adds local matching evidence to the global similarity path without requiring exhaustive pairwise matching. For Texas horned lizard, which lacks training labels, Stage A does not use this calibrated image-pair fusion path. It uses SIFT + LightGlue only in the uncalibrated cluster-level merge before the main correction in Stage C and Stage D.

3.1.3. Species-wise representation selection

During development, we evaluated frozen and fine-tuned global representations for Stage A. Fine-tuned variants were trained only when training identity labels were available, using ArcFace or Triplet loss depending on the species-specific setting [23, 24]. The Stage A representation for each species was selected according to Public Leaderboard ARI under the submitted development path. This keeps the final pipeline focused on the representation that best supported the initial clustering for each subset.

The selected representations reflect the data condition of each subset. Lynx and Salamander retain fine-tuned MiewID-msv3 variants. Lynx uses ArcFace fine-tuning, while Salamander uses a Triplet-loss fine-tuned representation. SeaTurtle keeps the frozen MegaDescriptor-L-384 representation. Texas horned lizard has no labeled training set, so supervised fine-tuning is not performed and the label-free fallback uses the frozen MiewID-msv3 and MegaDescriptor-DINOv2 similarity average. Table 7 summarizes the selected Stage A representations.

3.1.4. Calibration, HDBSCAN, and cluster-level local merge

The final part of Stage A combines score calibration, primary representation selection, HDBSCAN clustering, and one conservative cluster-level local merge. For species with training labels and local matching evidence, isotonic calibration maps heterogeneous score sources to a comparable similarity scale before fusion [7, 25]. This is needed because cosine similarity and SIFT + LightGlue match count have different numerical ranges, and directly combining them could cause the match count to dominate the fused score. Texas horned lizard does not use this calibrated local-fusion path. It uses an uncalibrated global fallback for HDBSCAN and then a conservative cluster-level SIFT + LightGlue merge.

HDBSCAN, a density-based clustering algorithm, is used because the number of test identities is unknown in the identity discovery and clustering setting [26]. Stage A constructs a species-wise pairwise similarity matrix and applies HDBSCAN to the corresponding precomputed distance matrix. For labeled species, this similarity matrix combines calibrated similarity from the selected primary global representation, DINOv2 similarity, and SIFT + LightGlue match counts on shortlisted image pairs. Texas horned lizard has no training labels, so it uses fixed HDBSCAN parameters on an uncalibrated MiewID-msv3 and MegaDescriptor-DINOv2 similarity average. Table 8 summarizes the primary representations and HDBSCAN parameters.

Table 7
Selected global representations used by Stage A for each species.

Species	Selected Stage A representation	Fine-tuning status
Lynx	MiewID-msv3, ArcFace fine-tuned	Fine-tuned
Salamander	MiewID-msv3, Triplet fine-tuned	Fine-tuned
SeaTurtle	MegaDescriptor-L-384	Frozen
Texas horned lizard	MiewID-msv3 + DINOv2 average	Frozen, no training labels

Table 8
Selected primary global representations and HDBSCAN parameters for Stage A. HDBSCAN is applied to a precomputed distance matrix derived from the species-wise similarity matrix.

Species	Primary global representation	Min. cluster size	Min. samples
Lynx	MiewID-msv3, ArcFace fine-tuned	15	1
Salamander	MiewID-msv3, Triplet fine-tuned	2	1
SeaTurtle	MegaDescriptor-L-384	4	1
Texas horned lizard	MiewID-msv3 + DINOv2 avg.	2	1

After HDBSCAN, images of the same individual may still remain split because of pose, camera angle, or local texture variation. Stage A therefore applies one conservative cluster-level local merge. Globally similar cluster pairs are shortlisted by centroid similarity and then verified by SIFT + LightGlue on representative images. A merge is accepted only when both the centroid-cosine threshold $\tau_{\cos}$ and the SIFT + LightGlue match-count threshold τ_{sift} are satisfied. For Texas horned lizard, the centroid shortlist uses frozen MegaDescriptor-L-384 features, while the HDBSCAN similarity matrix remains the uncalibrated MiewID-msv3 and DINOv2 average. Table 9 summarizes the thresholds and initial clustering statistics. These initial clusters serve as the clustering foundation for the later species-specific correction stages.

Table 9
Stage A conservative cluster-level local merge thresholds and initial clustering statistics.

Species	$\tau_{\cos}$	τ_{sift}	Test images	Initial clusters
Lynx	0.05	75	946	240
Salamander	0.10	75	689	456
SeaTurtle	0.10	20	500	158
Texas horned lizard	0.10	75	274	174

3.2. Stage B: Lynx orientation-aware merging

Stage B performs orientation-aware merging for Lynx. Its input is the Stage A cluster-label mapping, and its output is an updated mapping after merging cluster pairs that are likely to contain the same individual under different viewing orientations. This stage reduces orientation-induced under-merge, where left, right, front, or back views expose different local spots or body textures and cause Stage A to keep the same individual fragmented.

Stage B uses the provided orientation label to compare clusters at the orientation-group level. Images inside each cluster are first grouped by orientation label. Same-orientation groups, such as left/left or right/right, are compared directly with SIFT + LightGlue. Opposite lateral groups, such as left/right or right/left, use cross-view matching. The right-view image is horizontally flipped before matching so that the two side views are placed in a more comparable coordinate system. For each candidate cluster pair, the maximum match count over all compared image pairs is used as the cluster-pair local matching score.

A same-view candidate is merged when any compared image pair reaches a match count of at least 75, while a cross-view candidate uses a lower threshold of 50. The cross-view threshold is lower because opposite lateral views may not expose perfectly symmetric local textures even after flipping. To reduce cluster chaining risk, Stage B rejects any merge that would produce a cluster larger than 80 images. After Stage B, the number of Lynx clusters decreases from 240 to 218. The Public Leaderboard impact is reported in Sec. 4.

We further examined whether this cross-view (flip) path is justified for Lynx, since lynx flank pat-

terns are not obviously bilaterally symmetric. On the labeled Lynx training split, horizontally flipping the right-view image raises the mean same-individual SIFT + LightGlue match count from 1.2 to 3.3, but the same-individual versus different-individual separation is weak (ROC-AUC 0.557 over 600 same-individual and 600 different-individual left/right pairs). Lynx coat patterns therefore do not show the strong left/right facial-side correspondence reported for sea turtles [22]. However, at the operating threshold (match count $\geq$ 50), none of the 600 different-individual pairs qualified, so the cross-view path acts as a high-precision, low-recall bridge rather than relying on a symmetry assumption. Reproducing the submitted merge logic (Table 10), the cross-view path supplies only 4 of the 22 accepted merges and 88 of the same-individual image pairs recovered by Stage B, while same-view matching contributes the large majority (about 4,143 pairs). The flip is therefore a small but reliable addition, and same-orientation matching remains the main driver of Stage B.

Table 10
Stage B Lynx ablation: same-view-only versus adding the cross-view (horizontal-flip) path, reproducing the submitted merge logic. *Same-individual pairs* is the number of co-clustered Lynx test image pairs.

Variant	Merges (same/cross)	Lynx clusters	Same-individual pairs
Stage A (base)	– / –	240	11,590
Same-view only	20 / 0	220	15,733
Submitted (same + cross)	18 / 4	218	15,821

3.3. Stage C: Species-specific cluster-pair verification

Stage C performs conservative cluster-pair verification for Salamander, SeaTurtle, and Texas horned lizard. Its input is the cluster-label mapping after the previous applicable stages, and its output is an updated mapping after a small number of cluster merges. A merge is performed only when the corresponding species-specific rule is satisfied. Stage C is therefore a limited merge-based correction stage. Later stages still perform the main correction for Texas horned lizard and SeaTurtle.

3.3.1. Salamander

For Salamander, individual identity is strongly related to body markings and local texture. Stage C first retrieves same-orientation image candidates from the top 30 MiewID neighbors, keeping a pair when it is reciprocal within the top 30 or when its MiewID cosine similarity is at least 0.42. These image pairs are scored with SIFT + LightGlue and aggregated to cluster pairs. The submitted broad selector accepts a Salamander cluster pair only when all compared image pairs have the same orientation, the merged size and pair gain are bounded, and either a strong single image pair is present (match count $\geq$ 85 and MiewID cosine $\geq$ 0.30) or repeated moderate evidence is present (at least two edges with mean match count $\geq$ 55 and maximum match count $\geq$ 70). This rule is intentionally conservative because permissive merging can connect visually similar Salamander individuals under a few-shot identity distribution.

3.3.2. SeaTurtle

For SeaTurtle, Stage A leaves a mixed error pattern. Some same-individual images remain split, while visually similar individuals may already be incorrectly grouped. This subset is related to SeaTurtleID2022, a long-span sea turtle Re-ID dataset designed for reliable individual re-identification under realistic temporal variation [15]. Stage C therefore performs conservative positive merge verification using three evidence sources. First, it computes flip-aware global similarity by taking the maximum cosine similarity over original/original, original/flipped, flipped/original, and flipped/flipped image pairs. Second, it uses the bundled LightGlue match-count matrix to compare local facial scale patterns. Third, it builds four auxiliary HDBSCAN proposal partitions, including the leaf selection method with minimum cluster size $\in \{3, 4, 5\}$ and the excess-of-mass selection method with minimum cluster size set to 4, all with

minimum samples set to 1. A Stage C SeaTurtle merge requires support in all four proposal partitions, bounded merged size, sufficient flip-aware similarity, and sufficient local match evidence. The stage is still positive-only. The later signed relation repair in Stage F handles relation removals.

3.3.3. Texas horned lizard

For Texas horned lizard, Stage C uses local evidence from lizard foreground crops because the most useful individual cues are concentrated in ventral texture. This design is consistent with prior evidence that Texas horned lizards can be individually identified using ventral spot patterns [16]. Full images often contain the photographer's hand, ground background, and pose variation. Stage C uses SAM 3 to obtain mask-zeroed lizard foreground crops before local matching. We treat these foreground crops as belly-region crops because the images mainly expose the ventral side of the animal. Candidate cluster pairs are verified using ALIKED + LightGlue and SIFT + LightGlue on these crops [27]. The broad Stage C selector uses the same local matching conditions as Stage D, but applies them only as a capped merge rule. It accepts a pair when ALIKED ≥ 50 and SIFT ≥ 20, ALIKED ≥ 45 and SIFT ≥ 35, or SIFT ≥ 85, with a merged-size cap.

The effect of Stage C on Texas horned lizard is intentionally limited. Without training labels, the merge rules cannot be calibrated in the same way as for Lynx, Salamander, or SeaTurtle. Stage D later replaces the Texas horned lizard clustering with a new image-level graph and serves as the main correction stage for this species. Table 11 summarizes these cluster-count changes and their later handling.

Table 11
Cluster-count changes after Stage C and how each output is handled by later stages.

Species	Clusters before	Clusters after	Accepted merges	Later handling
Salamander	456	455	1	Retained for Stage E
SeaTurtle	158	145	13	Repaired in Stage F
Texas horned lizard	174	165	9	Replaced by Stage D

3.4. Stage D: Texas horned lizard belly-region graph replacement

For Texas horned lizard, Stage A is less reliable because no labeled training set is available for HDBSCAN parameter selection, similarity calibration, or train ARI-based primary feature selection. It therefore relies on uncalibrated global similarity and conservative local verification rather than on train-calibrated score fusion. At the same time, reliable identity cues are concentrated in ventral texture, while full images often include the photographer's hand, ground background, and pose variation. Stage D addresses this by replacing the Stage C Texas horned lizard result with a new image-level graph built from belly-region local matching.

The replacement is different from merge-based refinement. Stage C can connect clusters that were separated, but it cannot split different individuals that were already grouped together. Stage D instead discards the previous Texas horned lizard cluster units. Each image becomes a graph node, an edge is added when a belly-region local matching rule is satisfied, and connected components define the new clusters. This can recover reliable same-individual relations while reducing the direct inheritance of over-merged clusters from earlier stages.

Belly-region crop generation. Stage D applies SAM 3 to each Texas horned lizard image with "horned lizard" as the text prompt to obtain a foreground mask for the whole lizard [28]. The predicted mask score map is converted into a binary mask with threshold 0.5, every non-mask pixel is set to black, and the lizard body is cropped using the bounding box of the mask. SAM 3 does not further segment the anatomical belly region. We refer to these foreground crops as belly-region crops because

the images mainly expose the ventral side of the animal and the resulting crop mainly contains belly-region texture. Figure 2 illustrates this transformation from a full image with hand and background pixels to the foreground crop used for local matching.

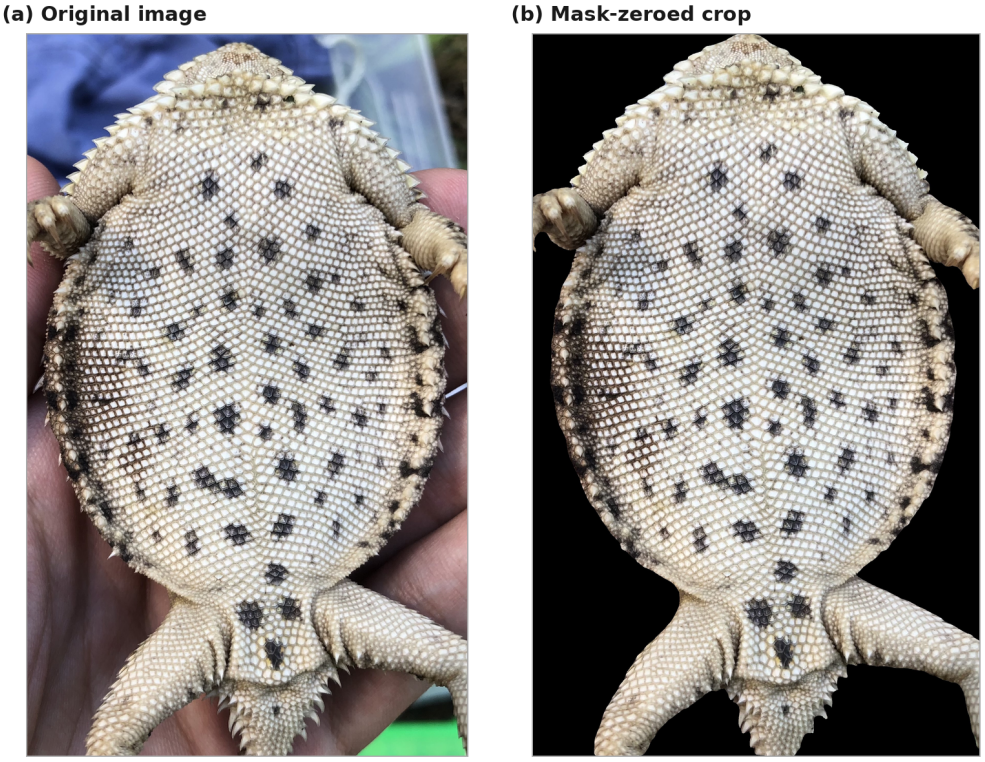

Figure 2: Example of Texas horned lizard foreground cropping for Stage D. (a) Original image with hand and background pixels. (b) SAM 3 mask-zeroed foreground crop. SAM 3 does not segment the anatomical belly region. The crop is used as a belly-region crop because the subset mainly exposes ventral texture.

Graph construction. Stage D applies ALIKED + LightGlue and SIFT + LightGlue to all pairs of the 274 belly-region crops, producing two aligned match-count matrices. Each crop is treated as an image-level node, and a candidate pair is added as an undirected edge only if at least one rule in Table 12 is satisfied. The first two rules require agreement between the ALIKED and SIFT matchers, while the third preserves cases where SIFT alone provides very strong local evidence. Figure 3 visualizes this edge decision for one accepted pair and one rejected pair under the two matchers. The accepted pair has strong support from both matchers and becomes an edge, whereas the rejected pair has insufficient local evidence and remains disconnected. The connected components of the accepted-edge graph become the new Texas horned lizard clusters.

Table 12
Texas horned lizard belly-region graph edge rules. Each crop is a node, and each accepted crop pair becomes an undirected edge.

Rule	Edge condition
1	ALIKED match count ≥ 50 and SIFT match count ≥ 20
2	ALIKED match count ≥ 45 and SIFT match count ≥ 35
3	SIFT match count ≥ 85

Stage D changes the Texas horned lizard result from 165 clusters to 213 clusters. Its main effect is therefore not to add many new merges, but to remove unreliable same-individual relations inherited

from Stage A and Stage C and rebuild a more conservative clustering from belly-region local matching evidence.

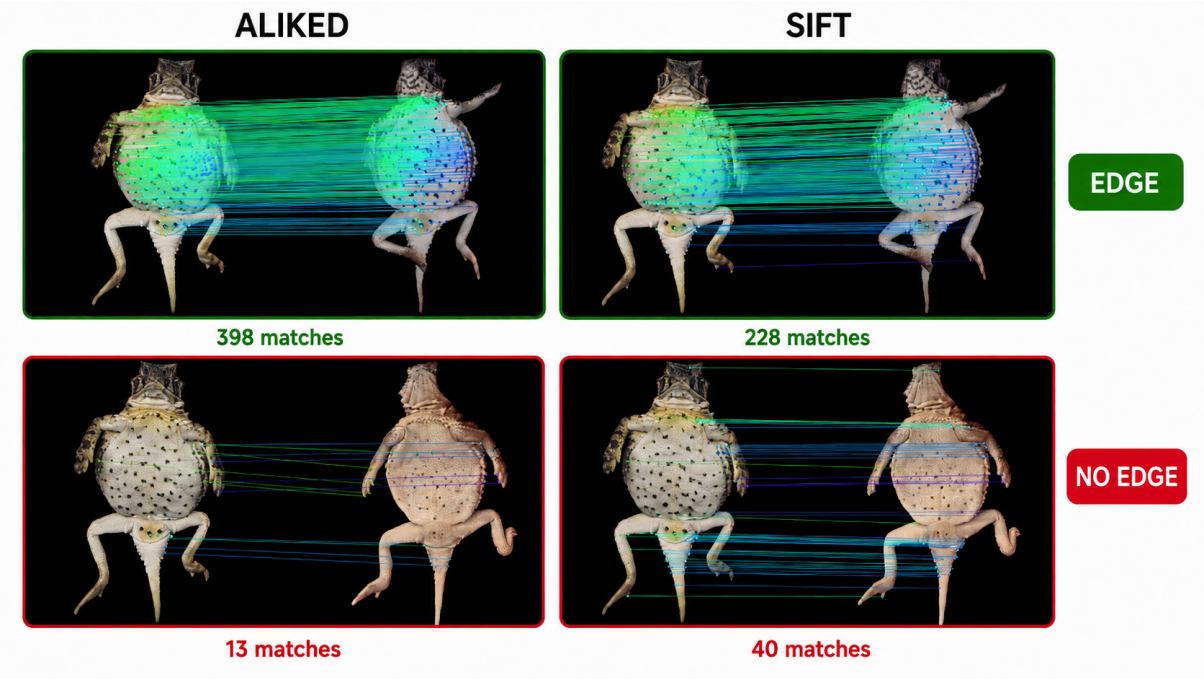

Figure 3: Example edge decisions in the Texas horned lizard belly-region graph. Columns show ALIKED + LightGlue and SIFT + LightGlue matches for each candidate crop pair. The top green row shows a pair that becomes an edge by Rule 1 with ALIKED = 398 and SIFT = 228. The bottom red row shows a rejected pair with ALIKED = 13 and SIFT = 40, satisfying no rule in Table 12.

3.5. Stage E: Salamander region-normalized merging

Stage E performs conservative region-normalized merging for Salamander. It is executed after Stage D. For Salamander rows (the rows of the cluster-label mapping whose images belong to the Salamander subset), this is equivalent to applying the rule on the Stage C mapping because Stage D changes only Texas horned lizard rows. Region-normalized merging means that merge evidence is recomputed on normalized body and head crop views rather than only on raw images. This is important because Salamander has a few-shot identity distribution, with 587 identities and 1,388 images, and broadly lowering the merge threshold can connect visually similar individuals into incorrect large clusters.

The most stable Salamander identity cues mainly come from yellow markings on the body and head, but raw images can be affected by background, hand occlusion, pose, and cropping differences. Stage E therefore uses a deterministic region extraction pipeline. It detects yellow-patch pixels in HSV color space, combines them with the surrounding dark body foreground to form a body mask, keeps the largest connected foreground components, square-pads the crop to produce a normalized body view, and rotates the body crop by its PCA-estimated major axis when that axis is reliable. The largest yellow-patch component inside the body crop is then used as the center for a head or main yellow-patch crop.

Stage E computes MiewID cosine similarity on three views, namely the raw image, body crop, and head crop. SIFT is applied only to the body and head crops, where identity-related markings are more concentrated. SIFT matches are filtered by the ratio test and a RANSAC-based affine consistency check, and the RANSAC inlier count is used as the local matching score [29]. The final merge rules are restricted to two reliable orientation branches, as shown in Table 13, and every accepted pair must also satisfy a pair-calibrator probability of at least 0.90, pair gain at most 16, and merged component size at most 12. After Stage E, the number of Salamander clusters decreases from 455 to 452, reflecting a small number of high-confidence region-normalized merges.

Table 13
Submitted Salamander region-normalized merge branches. Accepted pairs also satisfy the calibration, pair-gain, and size constraints described in the text.

View branch	Merge rule	Rationale
top/top	SIFT inliers ≥ 10 and raw MiewID cosine ≥ 0.75	top/top images already have relatively stable global similarity in the raw image view
right/right	SIFT inliers ≥ 15 and body MiewID cosine ≥ 0.60	right/right images rely more on body-crop matching, so a stricter SIFT inlier threshold is used

3.6. Stage F: SeaTurtle signed relation repair

The input to Stage F is the post-Stage-E mapping. For SeaTurtle rows (the rows of the cluster-label mapping whose images belong to the SeaTurtle subset), this is equivalent to the Stage C mapping because Stage D changes Texas horned lizard rows and Stage E changes Salamander rows. At this point, both SeaTurtle error directions may remain. Different individuals may be grouped together, while same-individual images may remain split because of camera angle, pose, or image quality. These difficulties are consistent with long-span sea turtle Re-ID, where temporal variation, viewpoint differences, and image quality can affect reliable individual matching. A merge-only step cannot repair clusters that already contain different individuals, and a split-only step may remove correct same-individual relations. Stage F therefore performs signed relation repair. A negative operation removes unreliable same-individual relations, and a positive operation adds high-confidence cross-cluster same-individual relations.

Negative repair. Stage F removes a same-individual relation only when independent image-evidence checks agree that it is visually unsupported. The first check recomputes local texture matching and global similarity for the image pair. The pair is treated as suspicious only when both scores fall below their thresholds. The second check uses an independently calibrated split signal, following the broader idea of supervised pairwise similarity verification in AnimalCLEF-style Re-ID systems [9, 30]. A relation is removed only when the independent split signals agree, reducing the risk of incorrect splitting.

Positive repair. Stage F then searches for image pairs in different clusters whose visual evidence suggests that they may belong to the same individual. To avoid new over-merge, each candidate pair must satisfy three conditions. The cosine similarity must be at least 0.45, the pair must form a reciprocal top-2 neighbor relation by similarity, and the connected component size after adding the relation must be at most 25. These constraints keep the relation-addition operation conservative.

After the negative and positive repair steps, Stage F rebuilds SeaTurtle clusters from the remaining and newly added same-individual relations. It is therefore not a one-way merge or split operation on existing clusters, but a relation-level graph update. The submitted release applies the frozen $r = 0.45$, rank-2, cap-25 repair map during inference and verifies that the pair-level changes match the frozen reference, with 74 same-individual pair relations added and 183 removed. After Stage F, the number of SeaTurtle clusters changes from 145 to 143. The improvement comes from removing unreliable relations and adding high-confidence cross-cluster same-individual relations.

4. Experimental Results

This section reports the development-path results of the proposed pipeline on the AnimalCLEF 2026 Public Leaderboard. The official metric is Adjusted Rand Index (ARI), and all reported scores are cumulative. Each later stage includes the output and corrections from previous stages. Because Public Leaderboard ARI was the main feedback signal during system development, these results should be interpreted as submitted development-path comparisons rather than independent causal ablations on a

separate held-out evaluation set. The final submitted release scored 0.74911 on the Public Leaderboard and 0.67537 on the Private Leaderboard.

We report two complementary views, cumulative stage-wise Public Leaderboard performance and species-level cluster-count changes. The first shows how the full pipeline changes after each stage, while the second helps interpret whether a stage mainly merges fragmented clusters, rebuilds a graph, or makes small relation-level adjustments.

4.1. Stage A component decomposition

Stage A reaches 0.58280 Public Leaderboard ARI, well above a global-descriptor baseline with density-based clustering (0.19401 in our reproduction; the official AnimalCLEF 2026 baseline is comparable, about 0.22). Because a natural question is whether this gain comes only from adding a local matcher, Table 14 decomposes the Stage A configuration along the submitted development path. The single largest component is indeed SIFT + LightGlue local matching with isotonic-calibrated fusion (+0.17382), but it reaches only 0.45294. The remaining gain to 0.58280 comes from square-padded inputs with train-ARI HDBSCAN parameter selection, a second DINOv2 global backbone, per-species fine-tuned representation selection, and the conservative cluster-level local merge. Stage A is therefore a combination of components rather than a single local-matching step, and the 0.45294 value in Table 6 corresponds to the local-matcher selection step of this decomposition.

Table 14
Stage A component decomposition along the submitted development path. Each row is cumulative and adds one component to the row above.

Cumulative Stage A configuration	ARI	Δ ARI
Global descriptors + density-based clustering (baseline)	0.19401	N/A
+ HDBSCAN with square-padded inputs	0.26444	+0.07043
+ train-ARI HDBSCAN parameter selection	0.27912	+0.01468
+ SIFT + LightGlue local matching, isotonic-calibrated fusion	0.45294	+0.17382
+ DINOv2 second global backbone	0.46297	+0.01003
+ per-species fine-tuned representation selection	0.50089	+0.03792
+ conservative cluster-level local merge (Stage A)	0.58280	+0.08191

4.2. Stage-wise Public Leaderboard performance

Table 15 summarizes the cumulative Public Leaderboard ARI along the submitted development path. Each row includes all preceding stages. Stage A reaches 0.58280 ARI as the shared clustering foundation. The later gains show that the remaining errors are species dependent. Stage B repairs Lynx orientation-induced fragmentation, Stage C adds only a small conservative gain, Stage D gives the largest gain by rebuilding Texas horned lizard from a belly-region image graph, and Stage E and Stage F provide smaller but consistent gains for Salamander and SeaTurtle.

The stage-wise trend supports the central design choice of this work. A shared global-local initial clustering stage is useful, but it is not sufficient for all species. Under the submitted development path, the +0.09748 ARI gain from Stage D suggests that Texas horned lizard is affected by unreliable full-image global clustering and inherited over-merge. For this subset, rebuilding the image-level graph from belly-region local evidence is more effective than applying small merge refinements to existing clusters.

Table 15
Cumulative Public Leaderboard ARI along the submitted development path. Each row includes all previous stages.

Stage	Correction added	ARI	Δ ARI
Stage A	Initial clustering	0.58280	N/A
Stage B	Lynx orientation merge	0.62184	+0.03904
Stage C	Three-species cluster-pair verification	0.62244	+0.00060
Stage D	Texas horned lizard belly-region graph replacement	0.71992	+0.09748
Stage E	Salamander region-normalized merge	0.72138	+0.00146
Stage F	SeaTurtle signed relation repair	0.74911	+0.02773

4.3. Per-stage correction footprint

Table 16 consolidates the concrete operation applied by each correction stage and how many same-individual relations it changed. The corrections are deliberately small and targeted: Stage C and Stage E accept only a handful of merges, Stage F adds 74 and removes 183 pairwise relations, and only Stage D rebuilds a species graph. This pattern matches the design intent that Stage A provides the shared clustering foundation while Stage B to Stage F apply conservative, species-specific corrections.

Table 16
Per-stage correction footprint. Each correction stage applies a targeted operation to a specific subset; Stage A (shared initial clustering) is omitted.

Stage	Species	Operation	Accepted changes
B	Lynx	orientation-aware merge	22 cluster merges (18 same-view, 4 cross-view)
C	Salamander, SeaTurtle, Texas horned lizard	cluster-pair verification	1, 13, and 9 merges respectively
D	Texas horned lizard	belly-region graph replacement	image-level graph rebuilt (165 → 213 clusters)
E	Salamander	region-normalized merge	3 merges
F	SeaTurtle	signed relation repair	74 relations added, 183 removed

4.4. Species-level cluster-count changes

Cluster count alone does not determine clustering quality, because ARI depends on pairwise consistency between the predicted clustering and the ground-truth clustering. However, cluster-count changes help interpret the structural role of each correction stage. Table 17 summarizes the number of clusters from Stage A to Stage F as structural diagnostics.

The cluster-count changes are consistent with the intended role of each correction stage. Lynx decreases from 240 to 218 clusters after orientation-aware merging. Salamander changes only from 456 to 452 clusters, reflecting the conservative design needed for a species with few images per individual. SeaTurtle decreases from 158 to 143 clusters, but Stage F is not merely a merge step because it modifies graph relations in both directions. Texas horned lizard provides the clearest example that fewer clusters are not necessarily better. Stage D increases the number of clusters from 165 to 213 while producing the largest Public Leaderboard ARI gain in the submitted development path. Overall, the final score comes from combining a shared initial clustering foundation with species-specific correction strategies, not from a single descriptor, threshold, or cluster-count target.

Table 17
Cluster counts after each stage. These are structural diagnostics, not direct quality scores. Only the stages that modify a species are shown; "–" marks a count carried forward unchanged from the previous shown stage.

Species	Stage A	Stage B	Stage C	Stage D	Stage E	Stage F
Lynx	240	218	–	–	–	–
Salamander	456	–	455	–	452	–
SeaTurtle	158	–	145	–	–	143
Texas horned lizard	174	–	165	213	–	–

5. Conclusion

This paper presented a two-phase, multi-stage clustering pipeline for the AnimalCLEF 2026 identity discovery and clustering task. The first phase builds a shared clustering foundation through Stage A by selecting species-wise global representations, constructing pairwise similarity matrices, applying HDBSCAN, and using local matching and isotonic calibration when training labels and local matching evidence are available. For Texas horned lizard, where no labeled training set is available, Stage A uses an uncalibrated global fallback and a conservative cluster-level local check. The second phase then applies species-specific correction through Stage B to Stage F according to the data condition and main difficulty of each species subset.

The submitted development-path results suggest that the shared initial clustering is useful but insufficient for this multi-species setting. Stage A reaches 0.58280 Public Leaderboard ARI, while the full pipeline reaches 0.74911 Public Leaderboard ARI and 0.67537 Private Leaderboard ARI after later correction stages. The improvements are not uniform across species. Lynx benefits from orientation-aware merging, Salamander requires conservative region-normalized merging, SeaTurtle requires signed relation repair for both over-merge and under-merge, and Texas horned lizard benefits most from belly-region graph replacement.

The largest cumulative gain comes from Stage D, which increases Public Leaderboard ARI from 0.62244 to 0.71992. Under the submitted development path, this suggests that small merge refinements are not sufficient when reliable identity cues are local and earlier clustering may contain inherited over-merge. Overall, the results support a practical design principle for multi-species Animal Re-ID clustering. A system should first build a shared initial clustering foundation and then apply species-specific correction rather than relying on one universal threshold or post-processing rule. Future work should evaluate these design choices on independent held-out data and further study belly-region crop quality for Texas horned lizard, relation-level repair for SeaTurtle, and region extraction stability for Salamander.

Acknowledgments

We thank the AnimalCLEF 2026 organizers, LifeCLEF, CLEF, and Kaggle for the challenge, evaluation platform, data release, identity annotations, and public model checkpoints. We also acknowledge the sea turtle and Texas horned lizard data teams, field experts, SeaTurtleID2022, the Texas horned lizard ventral spot-pattern study, WildlifeDatasets, and WildlifeReID-10k. We thank the Miin Wu School of Computing, National Cheng Kung University, for providing computational resources for this work.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI GPT-5.5 for grammar and spelling checks and for translation into English. The pipeline schematic in Figure 1 was designed and hand-drafted by the authors; OpenAI GPT-5.5 was used to re-render the authors' draft in a cleaner visual style. The

authors also used OpenAI GPT-5.5 image tools to arrange and annotate the matching panels in Figure 3; the matching panels themselves were produced by the implemented matching pipelines. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: AI challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[2] L. Adam, K. Papafitsoros, D. A. Williams, D. Biffi, L. Picek, Overview of AnimalCLEF 2026: Discovery and re-identification of individual animals, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[3] L. Adam, K. Papafitsoros, R. Kovář, V. Čermák, L. Picek, Overview of AnimalCLEF 2025: Recognizing individual animals in images, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 2898–2908. URL: https://ceur-ws.org/Vol-4038/paper_231.pdf.

[4] L. Otarashvili, T. Subramanian, J. Holmberg, J. J. Levenson, C. V. Stewart, Multispecies Animal Re-ID using a large community-curated dataset, 2024. doi:10.48550/arXiv.2412.05602. arXiv:2412.05602.

[5] V. Čermák, L. Picek, L. Adam, K. Papafitsoros, WildlifeDatasets: An open-source toolkit for animal re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5953–5963.

[6] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2025, pp. 2124–2134. doi:10.1109/CVPRW67362.2025.00197.

[7] V. Čermák, L. Picek, L. Adam, L. Neumann, J. Matas, WildFusion: Individual animal identification with calibrated similarity fusion, in: Computer Vision – ECCV 2024 Workshops, volume 15624 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 18–36. doi:10.1007/978-3-031-92387-6_2.

[8] R. Pakhomov, G. Demidov, K. Bogdan, S. Lanskikh, D. Dinmuhametov, A. Khlopotnykh, Individual wildlife recognition via hybrid global-local matching and segmentation-aware filtering, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3139–3146. URL: https://ceur-ws.org/Vol-4038/paper_251.pdf.

[9] N. Semenova, Meta-algorithm for open-set Animal Re-ID: WildFusion, XGBoost, and Dual-Backbone ArcFace, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3154–3165. URL: https://ceur-ws.org/Vol-4038/paper_253.pdf.

[10] D. Kim, B. Park, H. Bae, S. Lee, C. Lee, Fusion of global and local descriptors with feature calibration for multi-species animal re-identification: AnimalCLEF 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3069–3079. URL: https://ceur-ws.org/Vol-4038/paper_245.pdf.

[11] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[12] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, LightGlue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17627–17638. doi:10.1109/ICCV51070.2023.01616.

[13] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218. doi:10.1007/BF01908075.

[14] AnimalCLEF26 Organizers, AnimalCLEF26 @ CVPR & CLEF, https://www.kaggle.com/competitions/animal-clef-2026, 2026. Kaggle competition page and data release. Accessed: 2026-05-20.

[15] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7146–7156.

[16] D. Biffi, M. R. Tucker, A. Ackel, D. A. Williams, Identification of individual Texas Horned Lizards (Phrynosoma cornutum) using genotypes and ventral spot patterns, Ecology and Evolution 15 (2025) e71167. doi:10.1002/ece3.71167.

[17] L. Picek, J. Straka, M. Jiřík, E. Belotti, M. Duľa, J. Krausová, M. Bojda, V. Čermák, L. Bufka, R. Dvořák, L. Hrdý, V. Kocourek, J. Labuda, L. Toman, V. Trulík, M. Váňa, M. Kutal, CzechLynx: A dataset for individual identification and pose estimation of the Eurasian lynx, Scientific Data 13 (2026) 511. doi:10.1038/s41597-026-06853-9.

[18] Conservation X Labs, MiewID-msv3: Wildlife re-identification feature backbone, https://huggingface.co/conservationxlabs/miewid-msv3, 2025. Hugging Face model card. Accessed: 2026-05-20.

[19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research (TMLR) (2024). URL: https://openreview.net/forum?id=a68SUt6zFt.

[20] Bohemian Visual Recognition Alliance, MegaDescriptor-L-384: Animal re-identification feature backbone, https://huggingface.co/BVRA/MegaDescriptor-L-384, 2024. Hugging Face model card. Accessed: 2026-05-20.

[21] Bohemian Visual Recognition Alliance, MegaDescriptor-DINOv2-518: DINOv2-based animal re-identification feature backbone, https://huggingface.co/BVRA/MegaDescriptor-DINOv2-518, 2025. Hugging Face model card. Accessed: 2026-05-20.

[22] L. Adam, K. Papafitsoros, C. Jean, A. F. Rees, V. Čermák, Exploiting facial side similarities to improve AI-driven sea turtle photo-identification systems, Ecological Informatics 89 (2025) 103158. doi:10.1016/j.ecoinf.2025.103158.

[23] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699. doi:10.1109/CVPR.2019.00482.

[24] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[25] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 625–632. doi:10.1145/1102351.1102430.

[26] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, Journal of Open Source Software 2 (2017) 205. doi:10.21105/joss.00205.

[27] X. Zhao, X. Wu, W. Chen, P. C. Y. Chen, Q. Xu, Z. Li, ALIKED: A lighter keypoint and descriptor extraction network via deformable transformation, IEEE Transactions on Instrumentation and Measurement 72 (2023) 1–16. doi:10.1109/TIM.2023.3271000.

[28] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, C. Feichtenhofer, SAM 3: Segment anything with concepts, 2025. doi:10.48550/arXiv.2511.16719. arXiv:2511.16719.

[29] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395. doi:10.1145/358669.358692.

[30] S. Fedorchenko, S. Arefiev, Gradient boosting similarity in entity matching, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 3011–3018. URL: https://ceur-ws.org/Vol-4038/paper_240.pdf.